

WELFARE, CHILDREN, AND FAMILIES: A THREE-CITY STUDY
Wave 2, September 2000 – April 2001

USER'S GUIDE

Principal Investigators

Ronald J. Angel, University of Texas, Austin
Linda M. Burton, Pennsylvania State University
P. Lindsay Chase-Lansdale, Northwestern University
Andrew J. Cherlin, Johns Hopkins University
Robert A. Moffitt, Johns Hopkins University
William Julius Wilson, Harvard University

Funding Support

Federal Agencies: National Institute of Child Health and Human Development; Office of Disability, Aging, and Long-Term Care Policy, Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services; Administration on Developmental Disabilities, U.S. Department of Health and Human Services; Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services; Office of Research, Evaluation, and Statistics, Office of Policy, Social Security Administration; National Institute of Mental Health. Foundations: The Boston Foundation; The Annie E. Casey Foundation; The Edna McConnell Clark Foundation; The Lloyd A. Fry Foundation; Hogg Foundation for Mental Health; The Robert Wood Johnson Foundation; The Joyce Foundation; Henry J. Kaiser Family Foundation; W. K. Kellogg Foundation; Kronkosky Charitable Foundation; The John D. and Catherine T. MacArthur Foundation; Charles Stewart Mott Foundation; The David and Lucile Packard Foundation; Woods Fund of Chicago.

This volume documents the contents of survey data from *Welfare, Children and Families: A Three-City Study, Wave 2*. The data come from interviews conducted between September 2000 and June 2001 with 2,250 caregivers and 2,158 children in Boston, Chicago, and San Antonio.

Purpose of the study

The Welfare, Children and Families Study is a longitudinal study of children and their caregivers in low-income families that were living in low-income neighborhoods in three cities in 1999. The purpose of the study is to investigate the consequences of policy changes resulting from the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA)¹. The survey was designed to provide information on the health and cognitive, behavioral, and emotional development of children and on their primary caregivers' labor force behavior, welfare experiences, family lives, use of social service, health, and well-being. A detailed description of the research design can be found in *Welfare, Children and Families: A Three City Study, Overview and Design*, available at www.jhu.edu/~welfare or in hardcopy upon request.

Wave 2 Sample

We assume that the reader is familiar with the documentation from the first wave of this study, and we will not review all of the material presented there. See the design report described above or the Wave 1 User's Guide for a summary of how the sample was drawn and a description of each of the cities studied. This section describes the sample at wave 2 only.

At wave 1, the data file contained one record for each household, with data from the caregiver and focal child data on the same record. In contrast, at wave 2 we have *three* data files, two for caregivers and one for focal children. This is because not all caregiver/child pairs observed at wave 1 remained together at wave 2. Therefore, we followed to their new homes both caregivers and children who separated. This design enables users to study both caregivers and focal children longitudinally, without losing from the sample any dyad that separated between waves. In addition to tracking focal children, we interviewed children's new caregivers, using a modified version of the interview administered to continuing caregivers. This structure results in four types of respondents: continuing, separated, or new caregivers; and focal children. Continuing and new caregivers have been combined into a single file (N=2,187); separated caregivers are in another file (N=63); and data from focal children are in the third file (N=2,158). It is possible for one household from wave 1 to be represented on each of the three files: in the case that a caregiver and focal child have separated, that original household may be represented at wave 2 by a new caregiver interview, a separated caregiver interview, and a focal child interview.

The following table shows the pairings for focal children and caregivers at wave 2:

¹The survey is one component of a multidisciplinary project that also includes an embedded developmental study of about 600 children age 2 to 4 in 1999 and an ethnographic study of about 250 families residing in the same neighborhoods as the survey families.

Table 1. Focal child/caregiver interview pairings

<u>Interview Pair</u>	<u>N</u>
Focal child/continuing caregiver	2093
Focal child/new caregiver/separated caregiver	42
Focal child/new caregiver only	11
Focal child/separated caregiver only	6
Focal child only (no caregiver interview)	6
TOTAL	2158

Among caregivers, there are 38 continuing caregiver interviews, 11 separated caregiver interviews, and 3 new caregiver interviews that are not matched to a focal child interview.

Caregivers

Continuing and new caregivers

The wave 2 sample includes interviews with 2,131 continuing caregivers. These are women who were caring for the focal child at wave 1, and who continued to care for the child at the time of the wave 2 interview. In addition, the wave 2 sample includes interviews with 56 new caregivers.

We have combined data from continuing and new caregiver interviews into a single file for two reasons. First, this approach enables a data user to follow the child's living situation across waves. For example, from the perspective of a child who has changed households, the new caregiver's household income at wave 2 is more pertinent to the child's current situation than is the household income for the caregiver from whom the child is separated. Second, the continuing and new caregiver interview instruments each contain extensive data on the child's well-being and the caregiver-child relationship. The continuing/new caregiver data set includes observations on caregivers from 2,187 households.

The effective N is 2,182, because 5 cases interviewed at waves 1 and 2 were later determined to be duplicate interviews. These cases are flagged with the variable DUPLICAT. Because the data file has been in the public domain for several years with N=2,187, the Three-City Study team decided to preserve that sample size for the current release through ICPSR.

Features of the new caregiver interview

The new caregiver interview includes data on the caregiver's demographic characteristics. These data were not obtained a second time for continuing caregivers. The new caregiver also provides the date the focal child came to live with her.

Separated caregiver interview

The separated caregiver interview stands alone, and excludes all information about the child's well-being and the caregiver-child relationship. The separated caregiver provides the date that she and the focal child stopped living together. Combined with information from the new caregiver interview, a data user may assemble a month-by-month record of the child's place of residence between interviews. The data set includes observations on 63 caregivers.

Note that one respondent who was interviewed as a continuing caregiver appears to be separated from the focal child, based on her responses to some items. We have retained this case as a continuing caregiver interview, rather than moving it to the separated caregiver data set. Data users may decide how to treat this case. The value of the unique identifier HHID for this case is A0880030.

Child interview

Data from the child interview are included on a separate file from the caregiver interview, in contrast to the wave 1 file structure. The child interview is nearly identical to the wave 1 instrument, but includes a new section on Peer Association administered to all respondents age 10 or older.

At wave 1, sampled children were between the ages of 0 and 4 or 10 and 14. At wave 2, children are between the ages of 1 and 7 or 10 and 16.

Phone interview

Telephone interviews were administered to caregivers and/or children who had moved more than 100 miles away from the home where they were interviewed at wave 1. The telephone instrument was designed as a 45-minute interview, and is not as detailed as the in-person interviews. The variable CAPIMODE indicates whether the respondent participated in a telephone or in-person interview:

Table 2. Distribution of CAPIMODE

CAPIMODE	Continuing/new caregivers	Separated caregivers	Focal children
1=In-person interview	2150	62	2135
2=Telephone interview	37	1	23
TOTAL	2187	63	2158

The codebooks note those modules that were not administered to respondents who participated in telephone interviews. Within modules, some individual items were not administered. These skips have not been systematically documented in the codebooks, and users should use CAPIMODE to determine whether telephone respondents were excluded from the universe for a particular item.

Omitted cases

The public release version of the wave 1 data included 2,402 focal children and their caregivers. This file was edited to exclude 56 cases for whom the principal investigators determined data had been falsified. At wave 2, we successfully recontacted 45 of those households, resulting in interviews with 43 continuing caregivers, two new caregivers, one separated caregiver, and 40 focal children. These cases may be included in a cross-sectional analysis of the wave 2 data. However, any longitudinal analysis should exclude those cases, as there are no data available for them on the wave 1 file. A flag variable called NEWCASE indicates whether data on the respondent were included on the wave 1 file:

Table 3. Distribution of NEWCASE

NEWCASE	Continuing/new caregivers	Separated caregivers	Focal children
0=R is on wave 1 file	2142	62	2118
1=R is not on wave 1 file	45	1	40
TOTAL	2187	63	2158

At wave 2, we did not collect time-invariant demographic data or age data from continuing caregivers. Instead, we carried these data forward to the wave 2 file from wave 1. As a result, these data are missing on the wave 2 file for all previously omitted cases where the respondent is interviewed as a continuing caregiver.

Response rate

The wave 2 response rate is calculated as the percentage of respondents from wave 1 who provided partial or complete interviews at wave 2:

$$\text{Wave 2 response rate} = \frac{\text{\# of wave 2 respondents}}{\text{\# of wave 1 respondents}}$$

We use N=2,458 as the denominator. This number includes the 2,402 respondents on the wave 1 public release file, plus the 56 omitted cases we attempted to recontact at wave 2.

We report different response rates for children and for caregivers. For children, the response rate is 87.8 percent. For caregivers, we include all continuing and separated caregivers in the

numerator. At wave 2, there are 2,131 continuing caregivers, and 63 separated caregivers. The resulting response rate for that group is 89.2 percent. The response rate at wave 1 (the number of eligible respondents who gave a partial or complete interview) was 74.7 percent.²

Contents of the wave 2 instrument

Adult Portion

From adults, we gathered conventional measures of income, poverty, and family and labor force behavior B data that are generally useful for studies of the disadvantaged population. Specific questions address household structure, marriage, fertility, cohabitation, education, job history and characteristics, hours of work, earnings and wage rates, and sources of income. We also collect information on current and past welfare program participation, as well as participation in other programs such as Food Stamps, SSI, and so forth. In addition, we focus particular attention on the time respondents spend in welfare-related activities and information on actions related to job search. Wave 2 also includes a letter-word identification test for adult caregivers.

Child Portion

We focus on four main areas of child well-being: behavioral, cognitive, socio-emotional, and physical development. To assess these domains, the instrument combines measures used in large national studies with more detailed, process-oriented information on family functioning and child development using comprehensive measures. The questionnaire is designed to address environments and situations that pertain specifically to children in certain age groups while at the same time attempting to use similar measures across age groups in order to increase the longitudinal and cross-sectional comparability of findings. Throughout the child portion of the instrument, we have chosen measures that have proven validity and reliability in low-income and minority populations.

The survey instrument is composed of two computer-assisted personal interviews (CAPI). The first is a 100-minute interview conducted with the primary caregiver of the focal child. The second consists of standardized assessments of the child and a 30-minute interview, conducted if he or she is in the 10-16-year-old age group. Interview questions are organized into modules, each focusing on a different topic related to the lives of the children and caregivers in our study. The modules are listed in Table 3.

Items included in the available survey data fall into three categories: original items, recoded items, and constructed items. *Original items* are those asked of the respondent at the time of interview. *Recoded variables* are those that include changes to values in an original item. These changes are usually based on information from other items in the data set. One example from the welfare module is variable QWH58A21, a recoded version of survey item QWH58A, in which

² These response rates are the official rates, which were developed prior to the discovery that nine cases in wave 1 were duplicates (i.e., a second child from a previously interviewed household was included in the sample). Five of the nine cases were re-interviewed at wave 2. Adjusting the response rate formula to account for these cases results in a response rate for continuing and separated caregivers of $(2,126 + 63)/2,449 = 89.3$ percent, a difference of one-tenth of a percent compared to the original response rate. For focal children, the revised response rate is $(2,153)/2,449 = 87.9$ percent, also a difference of one-tenth of a percent compared to the original response rate.

original values were changed based on additional information from survey items QWH59A and QWH59B. *Constructed variables* combine information from several variables into a single item. In the Three-City Study data set, welfare receipt status (WELFST21) is one such item.

Table 4. Modules in the survey instrument

Caregiver interview

Demographics	Parenting Style* ⁺
Education and Training	Time Use*
Labor Force, Employment, and Work History	Father Involvement* ⁺
Self-Esteem/Self-Concept ⁺	Child Support* ⁺
Networks ⁺	Financial Strain Index ⁺
Housing	Welfare Participation and Experiences
Neighborhoods ⁺	Income
Family Routines Inventory* ⁺	Health and Disability
Home Environment* ⁺	Illegal Activities** ⁺
Positive Behaviors Scale* ⁺	Domestic Violence** ⁺
Child Behavior Checklist* ⁺	Brief Symptom Inventory**
Challenges to Parenting* ⁺	Woodcock-Johnson Letter-Word Identification ⁺

* Module excluded from separated caregiver interviews

** Administered by audio computer-assisted self interview (ACASI)

+ Module excluded from telephone interviews

Focal child interview

Physical measurements ⁺
Ages and Stages (ages 0-2) ⁺
Woodcock-Johnson (ages 4-16) ⁺
Letter-Word Identification
Applied Problems
Schooling (ages 10-16)
Peer Association (ages 10-16)
Child-Mother Relationship Scale* (ages 10-16)
Mother-Child Activities (age 10-16)*
Parental Monitoring* (ages 10-16)
Father Involvement* (ages 10-16)
Father-Child Relationship Scale* (ages 10-16)
Delinquency Scale* (ages 10-16)
Sex and Pregnancy* (ages 10-16)
Brief Symptom Inventory* (ages 10-16)

*Administered by audio computer-assisted self-administered interview (ACASI).

+Modules excluded from telephone interviews

Weighting

Because this survey is based on a clustered, stratified sample, we recommend that users employ weights in their statistical analyses. The survey contractor, RTI, Inc., developed household weights that take into account clustering, stratification, and non-response; and child-specific weights, which adjust further for the number of children in the household. The principal investigators developed “normalized” versions of the household and child weights that weight each city equally. (These are also referred to occasionally as “equalized” weights.) All four weights are included in the public release data set.

As of February 2007, the weights have been revised from their original values to correct for inconsistencies that have been identified during the course of data analysis. N=2,182 for all weights on the continuing/new caregiver file, N=63 on the separated caregiver file, and N=2153 on the focal child file; weight values are not included for duplicate cases (DUPLICAT=1). The weights included on the file have been trimmed so that weight values are top-coded at the 95th percentile. This step was taken to reduce the impact of outlying values on the weight variables.

RTI Household weights: (R3DUT5WT) These weights adjust the individual responses to account for the following factors:

Clustering: As noted above, we did not carry out a simple random sample, in which every household in a city would have had an equal probability of being selected. Rather, block groups were ranked according to percent poor for each city-specific racial/ethnic group; then a subset of block groups were randomly selected; and then households were randomly selected from within these block groups. The weights adjust for the probability that a given household was selected.

Stratification: A screening interview was conducted to see whether a family fell within one of the cells of the design matrix. Then families in specific cells were sampled at different rates in order to obtain a diverse sample. The weights adjust for the probability that a household in a given cell was selected.

Non-response: In addition, the weights were adjusted for the probability that a family responded to the screening interview and the main survey interview.

RTI Child weights: (R2CHT5WT) The child weights make one additional adjustment. Because RTI selected one child per household, children in large families were less likely to be chosen than children in small families. The child weight adjusts for the number of age-eligible children in the household. Use of the child weights allows the investigator to generalize to all children in the households selected.

Other things being equal, The RTI weighting procedure assigns larger values to households in cities with higher populations because these households had a smaller likelihood of being selected than did households in cities with lower populations. In this data set, respondents from Chicago

are weighted higher, all else equal, than respondents in San Antonio, who are in turn weighted higher than respondents in Boston because Chicago is the largest city and Boston is the smallest. So analyses that use the RTI weights will reflect information from Chicago more than information from the other cities, and information from San Antonio more than from Boston. If the investigator wants to report results that are proportional to population size, the RTI weights should be used.

Normalized weights: However, the principal investigators of the Three-City Study felt that this population was rather arbitrary and that it might be preferable to report results that weight each city equally and which therefore present the “average” experiences of households in the three cities. They modified the RTI household and child weights to create “normalized” (“equalized”) weights in which the total weights for households in one city equals the total weights in the other cities. The normalized household and child weights are called R2DUE5WT for dwelling units and R2CHE5WT for children. These *weights are recommended for data analysis*.

A note of caution: the normalized weights are applicable only if an analysis includes the entire sample. If a subset is used, that subset could be clustered in some of the cities and not others. And if so, a normalization performed for the whole sample will no longer weight each city’s selected households equally. Instead, the RTI weights would have to be normalized anew to preserve the equal-cities property. See Appendix A in the wave 1 user’s guide for a description of this procedure and sample syntax for the SAS system.

Should an investigator use the household weights or the child weights? There is no right or wrong answer to this question. The sampling design sampled children -- one per eligible household, and the inclusion of caregivers was ancillary to that choice. So in a statistical sampling sense, this is a sample of children. It could be argued, then, that the child weights are the better choice, and the principal investigators have used the child weights for nearly all of their analyses. However, if one is concerned solely about the adult caregivers, one might opt for the household weights; otherwise, the experience of caregivers with many children will be weighted more heavily than the experience of those with fewer children (because the child weights are larger for children with larger numbers of age-eligible siblings in their household).

Some survey analysts may wish to correct standard errors for weighting and clustering using statistical packages such as SUDAAN or STATA. After considerable analyses, the principal investigators concluded that adjustments for clustering had little effect on the size of the estimated standard errors. However, should you wish to adjust for clustering, you may use the variables and procedure described in the appendix in this user’s guide.

Unique Identifiers

HHID: Each household is assigned a household number that remains constant across waves and across interviews. This number appears as the household identifier HHID, a character variable. Use this variable to match households at wave 1 and wave 2, or to match respondents from the same household who appear on different data files (for example, caregiver and focal child at wave 2).

ZRID: Within each dataset, each observation is assigned a unique identifier that indicates from which household the interview is drawn, and in which interview the respondent is participating. ZRID is nearly identical to HHID, but it is one character longer, and the leading character indicates the interview type. Interview types are categorized as:

- 1: Wave 2 child interview (also wave 1 adult/child interview)
- 6: Continuing caregiver interview
- 7: New caregiver interview
- 8: Separated caregiver interview

An example:

Table 5. Unique identifiers

HHID	ZRID - wave 1 interview	ZRID - wave 2 continuing caregiver interview	ZRID - focal child interview
0180010	10180010	60180010	10180010

NEWID: Cross-wave caregiver-specific unique identifier. A household may have multiple associated caregivers over time. For example, a child may reside with her biological mother at wave 1, with her grandmother at wave 2, and again with her biological mother at wave 3. In order to facilitate tracking caregivers across waves, NEWID was created. Each caregiver at each wave is assigned a value on NEWID. NEWID takes the value HHID+01 for the wave 1 caregiver. Where a new caregiver is introduced, his/her values of NEWID is HHID+02. (After data cleaning, no cases were found where a child resided with one new caregiver at wave 2 and a different new caregiver at wave 3. Had that been the case, NEWID would equal HHID+03 for the second new caregiver.)

Recommendations to users

Delete 5 duplicate cases on the continuing/new caregiver file and focal child file for analysis, unless you are replicating an analysis published prior to February 2007. The 5 duplicate cases may be identified in either of two ways:

- 1.) Using the variable DUPLICAT, remove all cases where DUPLICAT=1
- 2.) Using any of the weight variables, remove all cases where a weight value is missing.

In a cross-wave analysis of caregivers, merge caregiver files by the variable NEWID. A household may have multiple associated caregivers over time. For example, a child may reside with her biological mother at wave 1, with her grandmother at wave 2, and again with her biological mother at wave 3. In order to facilitate tracking caregivers across waves, NEWID was created. Each caregiver at each wave is assigned a value on NEWID. NEWID takes the value HHID+01 for the wave 1 caregiver. Where a new caregiver is introduced, his/her values of NEWID is HHID+02. (After data cleaning, no cases were found where a child resided with one

new caregiver at wave 2 and a different new caregiver at wave 3. Had that been the case, NEWID would equal HHID+03 for the second new caregiver.)

Use normalized (“equalized”) weights for analysis. Re-normalize as necessary (following guidelines above).

Data users familiar with earlier versions of the data file may find the value on the variable QHHEX_2 (focal child’s gender) for one case has changed. In the course of cross-wave analysis, researchers on the Three-City Study team found inconsistent reports for the focal child’s gender. These cases were investigated, and the current values of QHHEX_2 are correct and consistent across waves.

Census geocodes

The following variables appear on the separated caregiver and continuing caregiver **restricted-use** data files:

<u>Variable name</u>	<u>Variable description</u>
GDTSFIPS	2000 Census FIPS Code, State
GDTCFIPS	2000 Census FIPS Code, County
GDTTR	2000 Census Tract ID
GDTBLKGR	2000 Census Block Group ID

APPENDIX TO USER'S GUIDE

Some survey analysts may wish to correct standard errors for weighting and clustering using statistical packages such as SUDAAN or Stata. This appendix proposes one method to prepare the data for such an analysis.

We have included the following variables on each data file (continuing/new caregiver, separated caregiver, and focal child):

SCRID: Screener ID

PU: Primary frame unit

SEGID: Segment identification number

SITE: City

1=Boston

2=Chicago

3=San Antonio

STR: Race/ethnicity stratum

B=Non-Hispanic Black/African-American

W=Non-Hispanic White

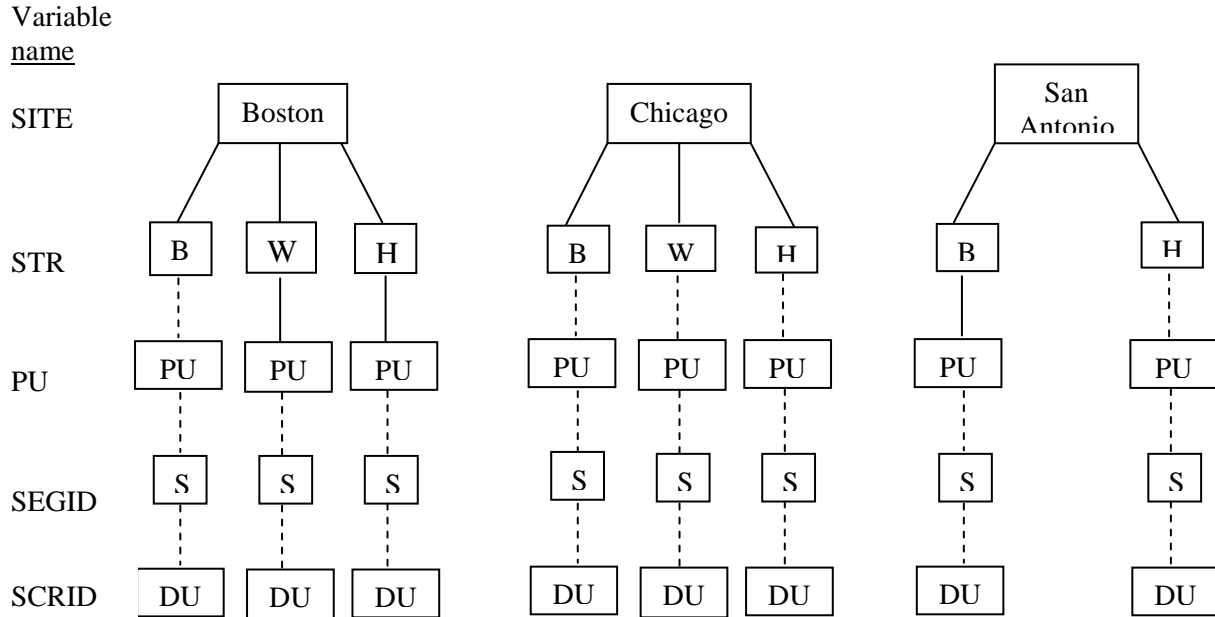
H=Hispanic/Latino

The firm conducting the survey, Research Triangle Institute (RTI), constructed eight sets of block groups from all of the blocks in the three cities in the 1990 Census. Each set ranked all the block groups in a city in descending order of the poverty rate of children in a particular race-ethnic group. In Boston, three such sets were compiled and ranked --one each for Non-Hispanic Whites, Non-Hispanic Blacks, and Hispanics. Three analogous sets were compiled for Chicago, and two sets for San Antonio--Non-Hispanic Blacks and Hispanics. Only block groups falling below a specified poverty level were retained in the sampling frame. The variable SITE refers to the city from which the block groups were drawn. The variable STR refers to the racial/ethnic composition of the block group.

The block groups in the sampling frame are referred to as primary frame units and are represented in the variable PU. In five of the eight sets described above, a random set of the PUs was selected with probability proportional to size. These five sets are referred to as "non-certainty strata," meaning that not all of the eligible PUs were selected into the sample. In the other three sets, all of the PUs were selected. These are certainty strata. Within these strata, all values of the variable PU are blank.

From each stratum, the selected PUs were divided into segments, which are areas of a size typically regarded as convenient for surveying and generally consist of 90-120 dwelling units. A set of segments was chosen randomly from the selected PUs. All selected segments were then counted and listed (i.e., interviewers visited the segments, counted the housing units, and wrote down the addresses of all occupied dwelling units). Segments are represented in the variable SEGID. A random sample of dwelling units (identified by street addresses) was then selected from within each segment. These dwelling units are the households that appear in our sample. The unique identifier SCRID represents each household.

The schematic below shows how households were selected into the sample. The solid lines represent points at which all eligible units were included. The dashed lines represent points where only a subset of units (whether PU's, segments, or dwelling units) was selected.



The method proposed below to account for clustering was developed by the survey contractor. Different methods are used for households in certainty and non-certainty strata. Data users may wish to use other techniques with which they are familiar. For a general discussion of data preparation for complex survey data analysis, see Eltinge and Sribney (1996), Chapter 16 in Levy and Lemeshow (1999), and StataCorp (2003).

For certainty strata (White and Hispanic strata in Boston (site=1, str= "W" or "H") and the Black stratum in San Antonio (site=3, str= "B")):

Sort all households by SITE, STR, SEGID, SCRID. This sorts the data into a geographically ordered list, with households ordered by segment within a site/race-ethnic group. The variable PU takes no value for these cases. Go down the list and form pairs of dwelling units in a new variable called STRATA. Number these pairs from 1 to N/2. Then call each of the dwelling units within a stratum a CLUSTER. The clusters will take on the values of 1 or 2.

For example, assume you are working with a sample that includes 1000 cases from the certainty strata. The first two observations will have the value 1 on the new variable STRATA, the next two observations will have the value 2, and so on. The last two cases will have the value 500. Within each pair, the first observation will have the value 1 on the new variable CLUSTER. The second observation will have the value 2. A list of these data would have the following appearance:

SCRID	SITE	STR	PU	SEGID	STRATA	CLUSTER
1201010S	1	H	.	1201	1	1
1201020S	1	H	.	1201	1	2
1201030S	1	H	.	1201	2	1
1201040S	1	H	.	1201	2	2
1202010S	1	H	.	1202	3	1
1202020S	1	H	.	1202	3	2
.						
.						
.						
2561010S	3	B	.	2561	500	1
2561020S	3	B	.	2561	500	2

No pair should cross between two site/race-ethnic groups. If there is an odd number of dwelling units in an area, the last dwelling unit should be assigned a cluster value of 2 and attached to the last pair in its site/race-ethnic group, so that the last “pair” would actually include three observations.

For non-certainty strata:

Sort all households by SITE, STR, SEGID, and PU. Here, definitions of clusters and strata are not based on the dwelling units and dwelling unit pairs within the non-certainty strata. Rather, PUs and PU pairs are used for cluster and strata definition. Again, the pairing of PUs should be done within a common area, so that pairs do not cross area type.

For example, assume you are working with a sample that includes 1000 observations from the non-certainty strata. Among those 1000 observations, 200 PUs are represented. The number of PU pairs that would emerge from this sample would be (# of PUs)/2, or 200/2=100. All of the observations within the first PU in a given pair would carry a value of 1 on the variable CLUSTER. The observations within the second PU would carry a value of 2. A list of these data would have the following appearance:

SCRID	SITE	STR	PU	SEGID	STRATA	CLUSTER
1401010S	1	B	1	1401	1	1
1401020S	1	B	1	1401	1	1
1401030S	1	B	1	1401	1	1
1401040S	1	B	1	1401	1	1
1402010S	1	B	5	1402	1	2
1402020S	1	B	5	1402	1	2
1403010S	1	B	7	1403	2	1
1403020S	1	B	7	1403	2	1
1403030S	1	B	7	1403	2	1
1404010S	1	B	9	1404	2	1
1404020S	1	B	9	1404	2	2
1404030S	1	B	9	1404	2	2

```

.
.
2261010S    3      H    140    2261        100        2
2261020S    3      H    140    2261        100        2

```

In Stata's svyset command, the analyst may set the variable STRATA as the strata identifier variable, and the variable CLUSTER as the PSU (cluster) identifier variable.

References

Eltinge, J.L. and W.M. Sibney (1996). svy3: Describing survey data: sampling design and missing data. *Stata Technical Bulletin* 31: 23-26. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 235-239.

Levy, Paul S. and Stanley Lemeshow (1999). *Sampling of Populations: Methods and Applications*. 3rd ed. New York: John Wiley & Sons, Inc.

StataCorp (2003). *Stata Survey Data Reference Manual, Release 8*. College Station, TX: Stata Press.